Applied Data Science Capstone

IBM Data Science

Ameya Koranne
04/10/2023

IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ameya Koranne
04/10/2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- **Summary of methodologies:**
  - Data Collection API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis (EDA) Using SQL
  - EDA with Data Visualisation
  - Interactive Visual Analytics with Folium
  - Interactive Dashboard with Plotly Dash
  - Machine Learning Predictive Analysis (Classification Method)
- **Summary of all results:**
  - Exploratory Data Analysis Results
  - Interactive Analytics and Dashboard (Screenshots)
  - Predictive Analysis Results

Introduction

- **Project background and context:**
  SpaceX advertises Falcon 9 rocket launches on its website with a cost of $62-Million unlike other providers which cost upwards of $165-Million each. Much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch by gathering information about SpaceX and creating dashboards for the team. We will also determine if SpaceX will reuse the first stage. We will train a machine learning model and use public information to predict if SpaceX will reuse the first stage.
  We can then help SpaceY to compete with SpaceX with our insights.

- **Problems you want to find answers:**
  - How to improve the chances of first stage successful landings?
  - What factors affect the first stage successful landings?
  - What is the probability of using first stage again after successful landing?

Section 1

Methodology

Methodology

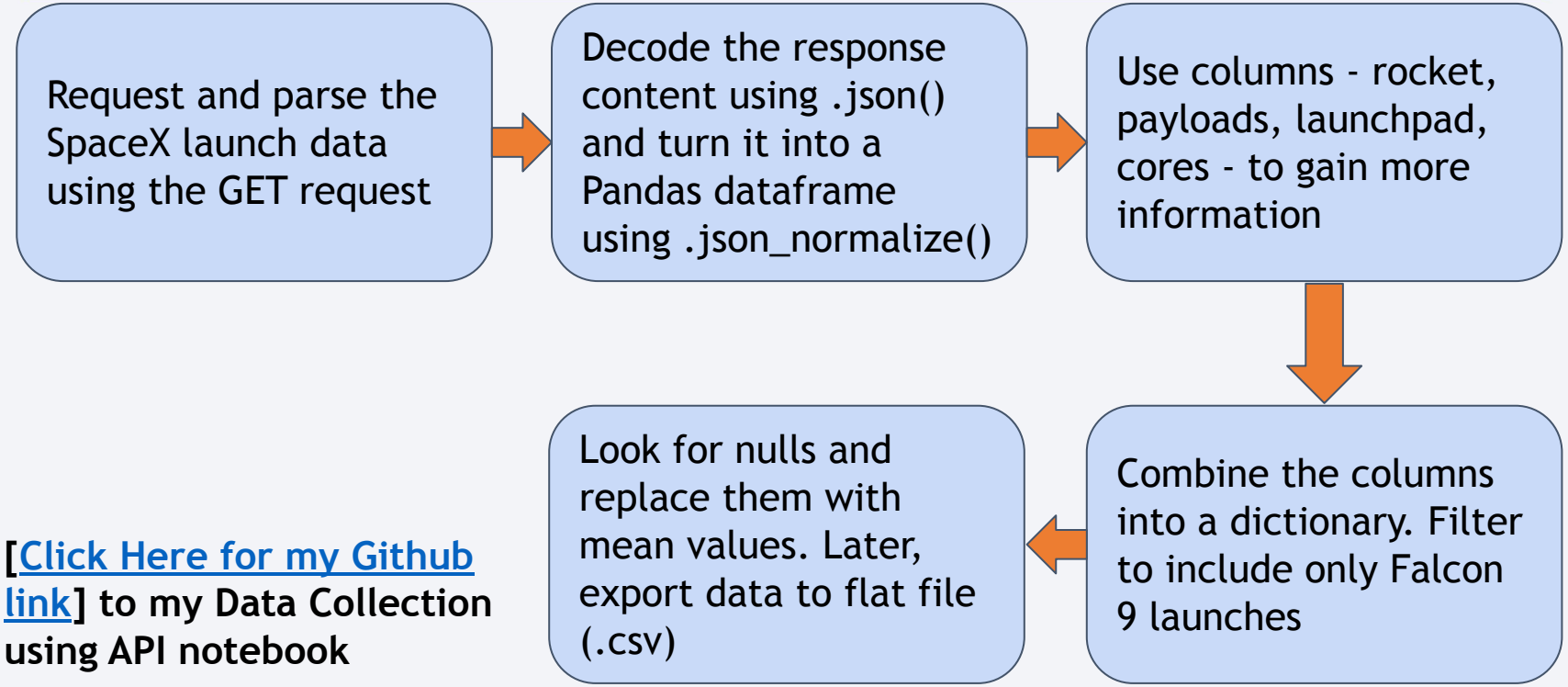
Executive Summary

- Data collection methodology:
  - Collected using SpaceX Rest API
  - Collected by Web Scraping from Wikipedia
- Performed data wrangling:
  - Data filtering
  - Replacing null values with mean values where applicable
  - One Hot Encoding to turn categorical values into numerical values
- Performed Exploratory Data Analysis (EDA) using visualisation and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models:
  - Logistic Regression, Support Vector Machine, K-Nearest Neighbours, Decision Tree were built, tuned and evaluated to determine best result.

Data Collection

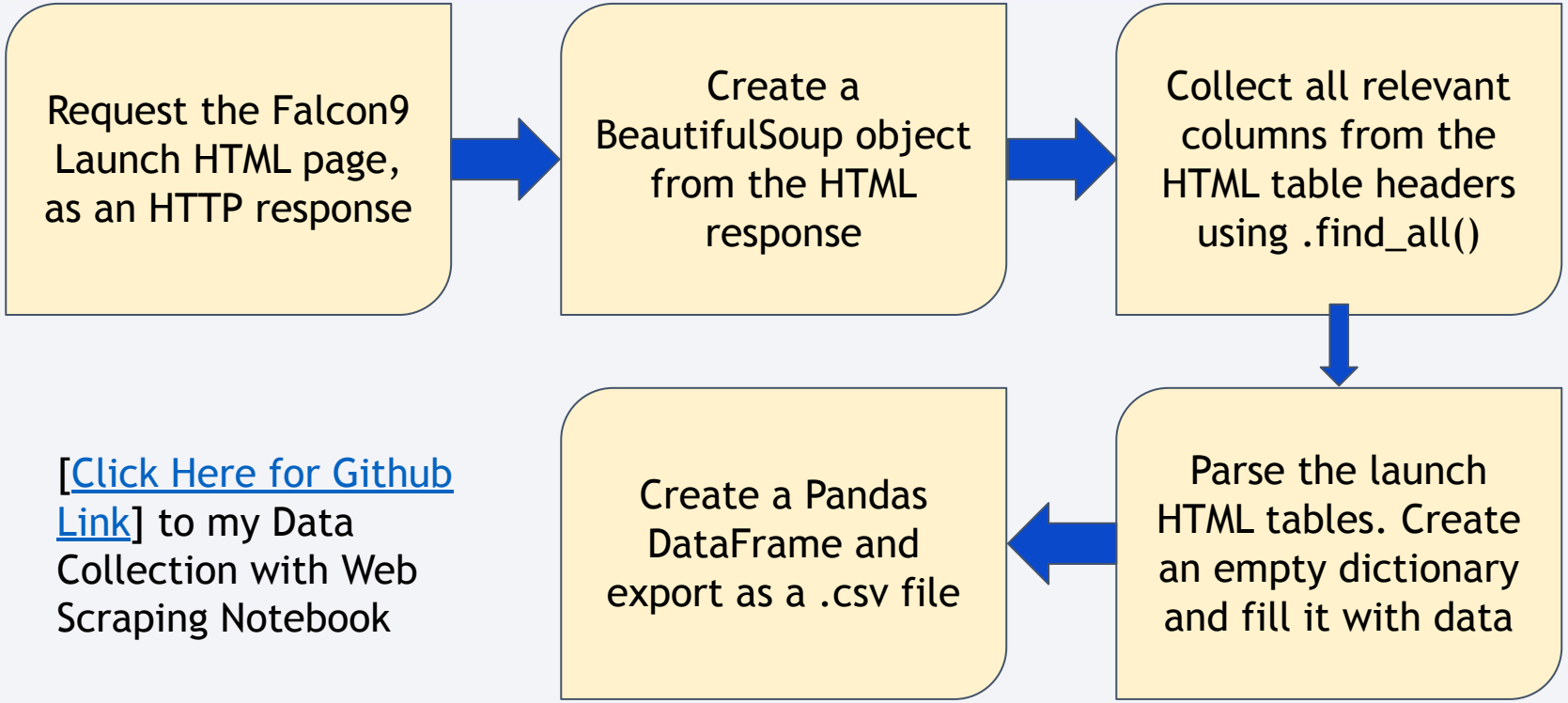
- Data columns collected using Spacex Rest API include:
  - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, Launchsite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- However, we need more information such as FlightNo., Time, etc which we collected from Wikipedia using web scraping. The following columns were collected:
  - Flight No., Date and time, Launch site, Payload, Payload mass, Orbit, Customer, Launch outcome
- The collected data was stored in a Pandas DataFrame, processed and exported to a .csv file

Data Collection – SpaceX API



[Click Here for my Github link] to my Data Collection using API notebook

Data Collection - Scraping



[Click Here for Github Link] to my Data Collection with Web Scraping Notebook

Data Wrangling

- We will start with converting outcomes with labels - True Ocean, True RLTS, True ASDS - into Training Labels with “1” meaning the booster successfully landed

- And outcomes with labels - False Ocean, False RLTS, False ASDS - to “0” meaning it was unsuccessful

- Perform calculations to learn:
  - Number of launches on each site
  - Occurences of each orbit
  - Number of landing outcomes
  - Export to .csv file

[Click Here for Github Link] to my Data Wrangling Notebook

EDA with Data Visualization

- Scatter Plots help us understand the relationship between two continuous variables. They are great with complex data. So we used scatter plots to chart variables to learn the relationship between them:
  - FlightNumber vs. PayloadMass
  - FlightNumber vs. LaunchSite
  - PayloadMass vs. LaunchSite
  - FlightNumber vs. Orbit
  - PayloadMass vs. Orbit
- In addition to scatter plots we used a bar chart to compare success rate for each orbit and a line chart to view the timeline of average success rate.

[Click here for Github Link] to my EDA with Data Visualisation Notebook

EDA with SQL

- **Summary of SQL queries we performed:**
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was achieved
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass (Used a subquery)
  - List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

- [Click Here for Github Link] to my EDA with SQL Notebook

Build an Interactive Map with Folium

- We added markers with circle, pop-up label, and text label of all Launch Sites using their coordinates (latitude and longitude) to show the respective geographical locations on the interactive map we built with Folium
- We added coloured markers (green for successful launches, red for failed launches) for each launch site on the map using Marker Cluster object to easily identify which launch sites have relatively high success rates
- We added a line marker to calculate the distance from closest highway, railway, coastline, and city from a launch site.

[Click Here for Github Link] to my Interactive Map with Folium Notebook

Build a Dashboard with Plotly Dash

- We added a dropdown list for all Launch Sites (to select either a specific launch site or to select combined launch sites)
- Selecting a specific launch site or all launch sites option, we can view the pie chart which shows the success rate vs. failure rate
- We added a Range Slider to Payload Mass variable with selectable points at 0, 2500, 5000, 7500, 10000
- Added a scatter plot which operates on above range slider for payload mass and launch success.
- [Click Here for Github Link] to my Dashboard with Plotly Dash file

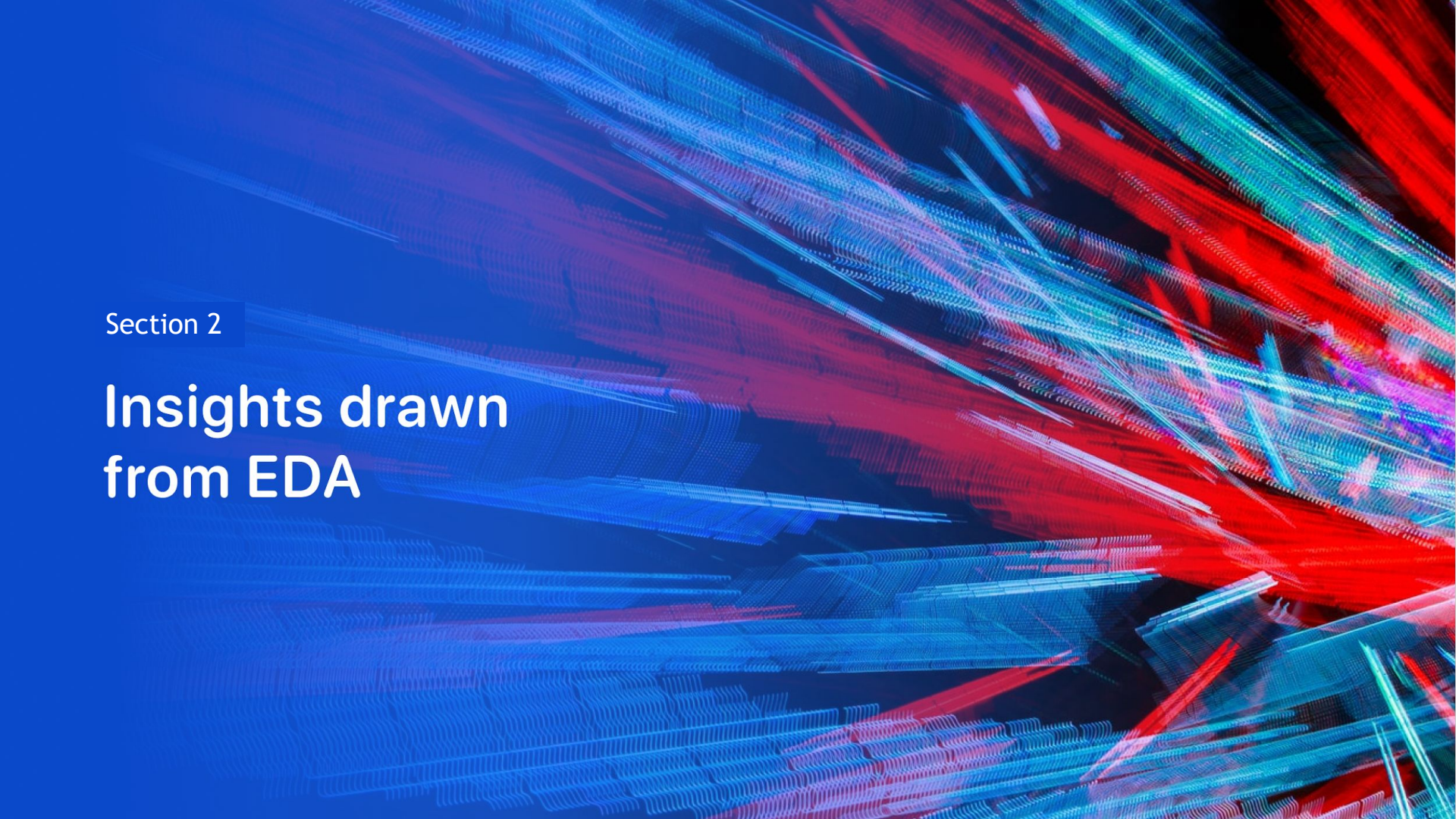
Predictive Analysis (Classification)

- We created a NumPy array based on column ‘Class’ in data
- We standardised the data using StandardScaler() and then transformed the data. We also split our data using train_test_split()
- We created GridSearchCV object to find the best parameters and calculate the accuracy of test data by using method ‘score’
- We used Logistic Regression, Support Vector Machine, K-Nearest Neighbour, and Decision Tree
- We found which method performs best

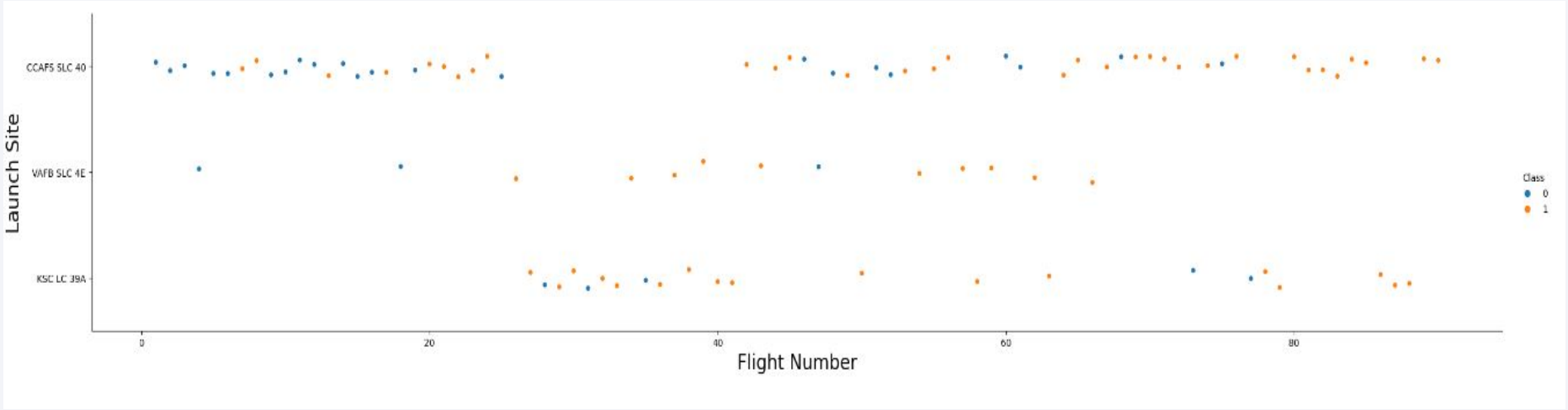
[Click Here for Github Link] to my Predictive Analysis Notebook

Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

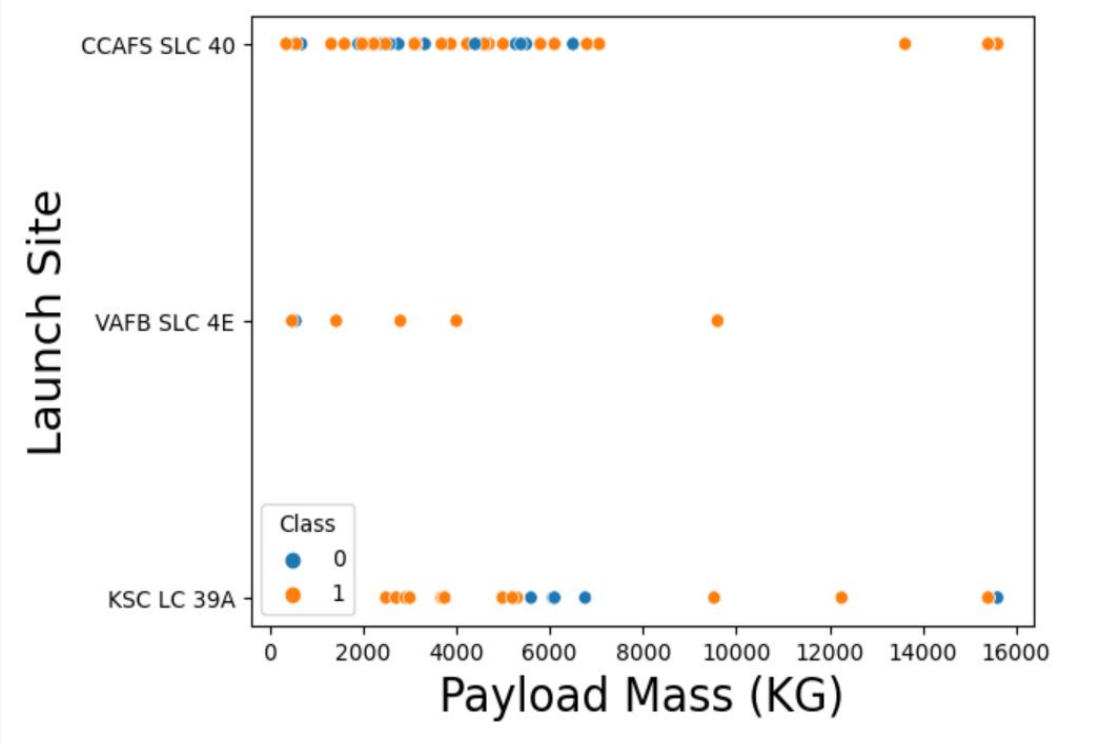


Observations:

- Launch Site CCAFS SLC 40 has highest number of launches
- VAFB SLC 4E and KSC LC 39A have lesser number of launches but their success rates are higher
- Success Rate is improving for all sites as the number of launches grow

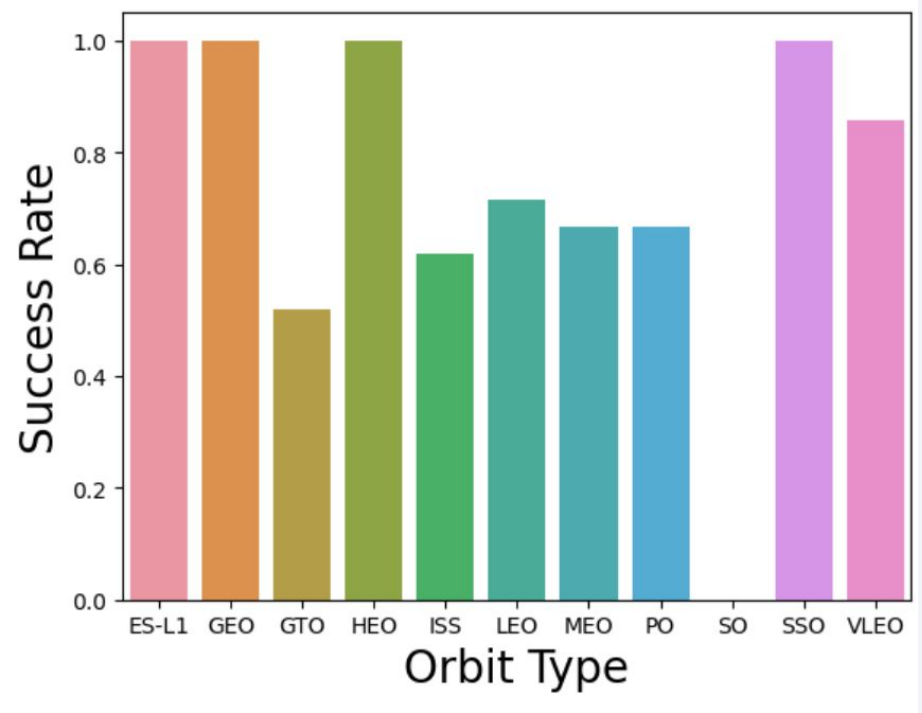
Payload vs. Launch Site

- For VAFB SLC 4E launch site, there are no rockets launched for heavy payload mass(greater than 10000 Kgs)
- CCAFS SLC 40 excels in heavy payload mass rocket launches with no failure so far
- KSC LC 39A has one failed launch at payload mass of almost 16000 Kgs



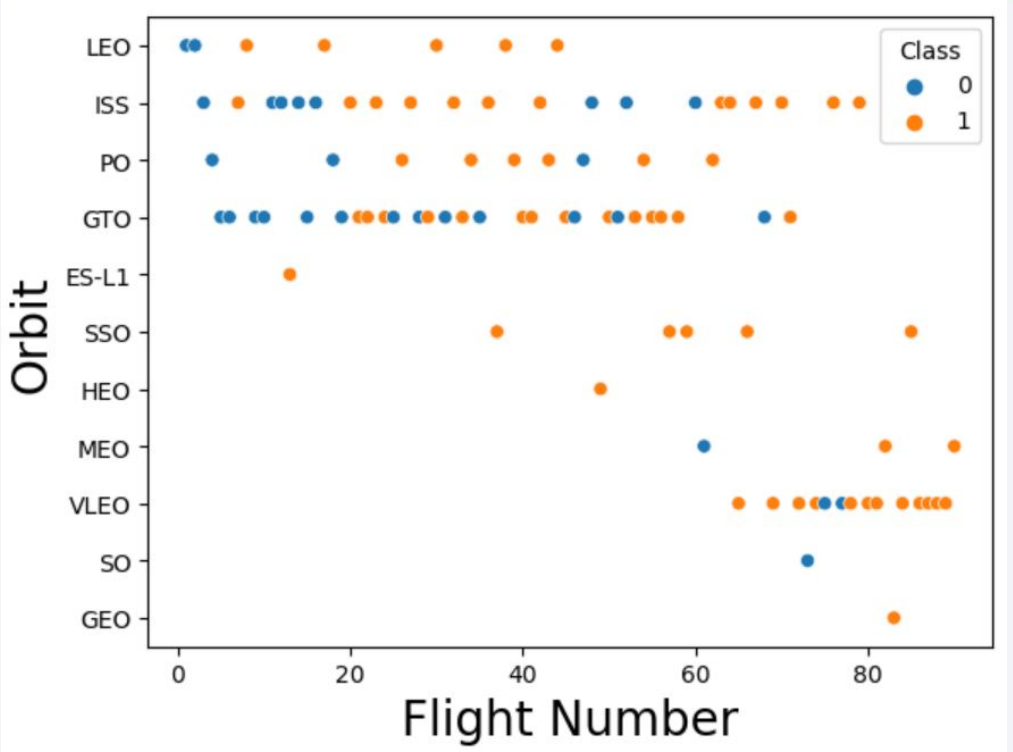
Success Rate vs. Orbit Type

- Orbits ES-L1, GEO, HEO, and SSO have a perfect status, they have not failed.
- Orbit ‘SO’ has a success rate of 0%
- Rest of the orbits have success rate between 50% and 85%



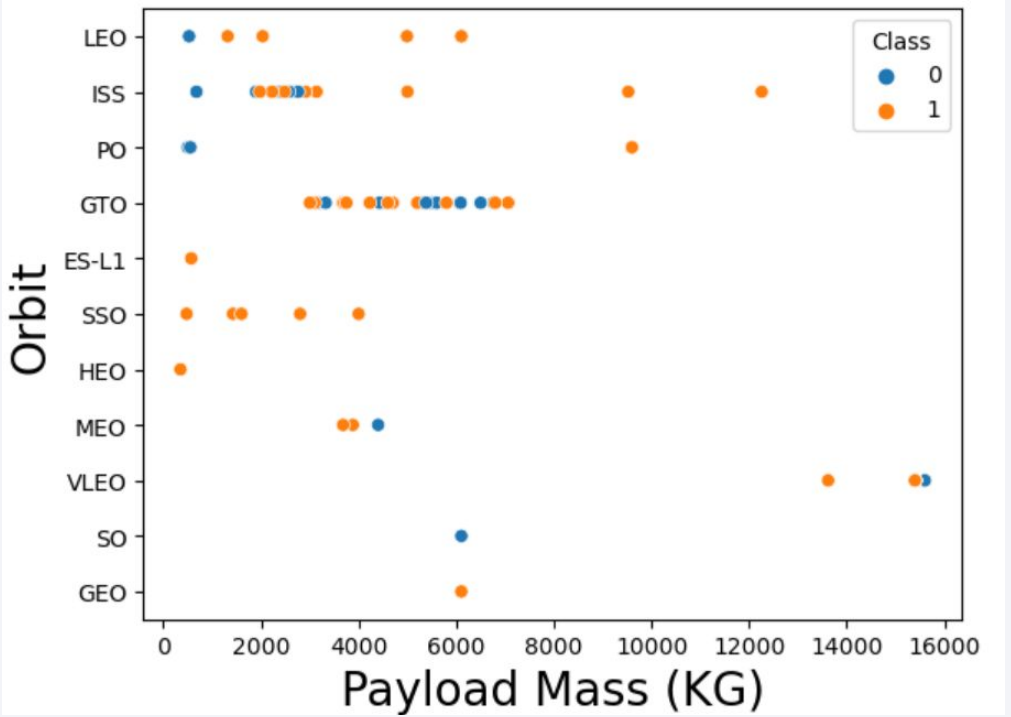
Flight Number vs. Orbit Type

- In VLEO orbit the success appears related to the number of flights
- ES-L1, SO, and GEO are the least preferred orbit type
- There seems to be no relationship between flight number when in GTO orbit



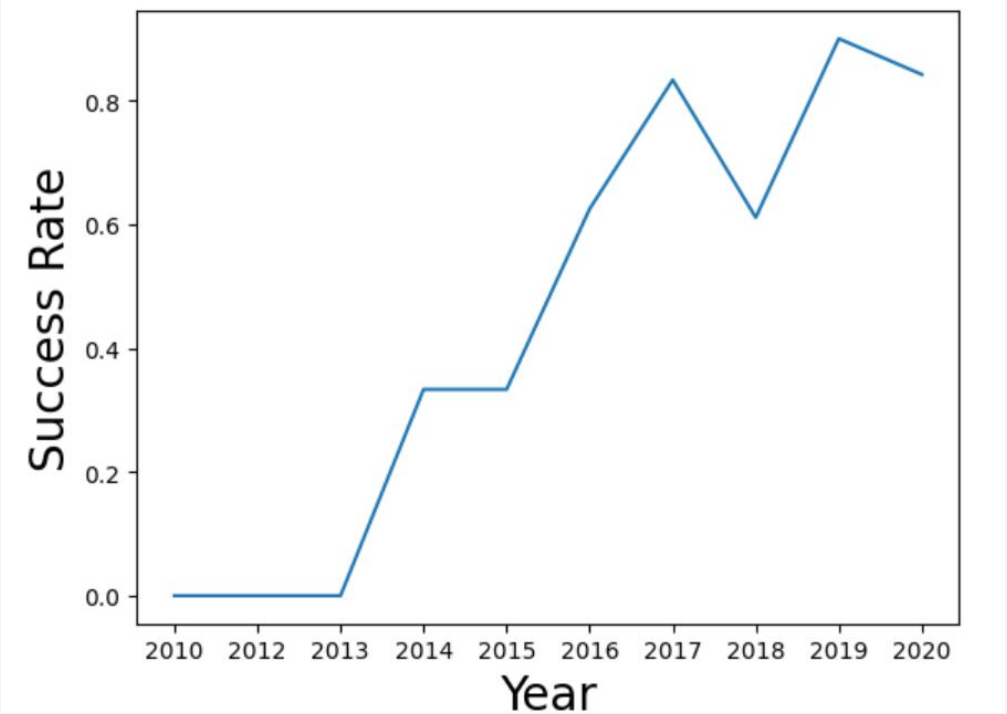
Payload vs. Orbit Type

- With heavy payloads the successful landing rates are more for PO, LEO and ISS
- However for GTO we cannot distinguish this as both positive and unsuccessful landings are both present



Launch Success Yearly Trend

- We can observe that the success rate from 2013 kept increasing till 2020



All Launch Site Names

- Find the names of the unique launch sites

Display the names of the unique launch sites in the space mission

In [35]:
```
%sql SELECT DISTINCT(LAUNCH_SITE) FROM SPACEXTBL
```

```
 * ibm_db_sa://ktk71166:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/BLUDB
Done.
```

Out[35]:

launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

Task 2

Display 5 records where launch sites begin with the string 'CCA'

In [36]:
```
%sql SELECT LAUNCH_SITE FROM SPACEXTBL WHERE (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5;
```

```
 * ibm_db_sa://ktk71166:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/BLUDB
Done.
```

Out[36]:

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- Calculate the total payload carried by boosters from NASA(CRS)

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

In [53]:
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) as total_payload_mass_kg\
    FROM SPACEXTBL \
    WHERE CUSTOMER = 'NASA (CRS)';
```

```
 * ibm_db_sa://ktk71166:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/BLUDB
Done.
```

Out[53]:

total_payload_mass_kg
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

In [54]:
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) \
    FROM SPACEXTBL \
    WHERE BOOSTER_VERSION = 'F9 v1.1';
```

```
 * ibm_db_sa://ktk71166:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/BLUDB
Done.
```

Out[54]:

1
2928

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

In [64]:

```
%sql SELECT MIN(DATE) \
FROM SPACEXTBL \
WHERE LANDING_OUTCOME = 'Success (ground pad)'
```

```
 * ibm_db_sa://ktk71166:***@55fbc997-9266-4331-afd3-
Done.
```

Out[64]:

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

In [70]:

```
%sql SELECT PAYLOAD, BOOSTER_VERSION, PAYLOAD_MASS__KG_ \
FROM SPACEXTBL \
WHERE LANDING_OUTCOME = 'Success (drone ship)' \
AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

```
 * ibm_db_sa://ktk71166:***@55fbc997-9266-4331-afd3-888b05e734c0.
Done.
```

Out[70]:

payload	booster_version	payload_mass__kg_
JCSAT-14	F9 FT B1022	4696
JCSAT-16	F9 FT B1026	4600
SES-10	F9 FT B1021.2	5300
SES-11 / EchoStar 105	F9 FT B1031.2	5200

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

In [71]:
```
%sql SELECT MISSION_OUTCOME, COUNT(*) as total_number \
FROM SPACEXTBL \
GROUP BY MISSION_OUTCOME;
```

```
 * ibm_db_sa://ktk71166:***@55fbc997-9266-4331-afd3-888b05e734c0.
Done.
```

Out[71]:

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

In [74]:
```
%sql SELECT BOOSTER_VERSION \
FROM SPACEXTBL \
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

```
 * ibm_db_sa://ktk71166:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.d
Done.
```

Out[74]:

booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

(Used a subquery)

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

In [82]:
```
%sql SELECT substr(Date,6,2) as month, DATE, BOOSTER_VERSION, LAUNCH_SITE, (LANDING_OUTCOME) \
FROM SPACEXTBL \
where (LANDING_OUTCOME) = 'Failure (drone ship)' and substr(Date,1,4)='2015';
```

```
 * ibm_db_sa://ktk71166:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/BLUDB
Done.
```

Out[82]:

MONTH	DATE	booster_version	launch_site	landing_outcome
10	2015-10-01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

In [93]:

```
%sql SELECT (LANDING_OUTCOME), count(*) as count_outcomes \
FROM SPACEXTBL \
WHERE DATE between '20100604' and '20170320' group by (LANDING_OUTCOME) order by count_outcomes DESC;
```

```
 * ibm_db_sa://ktk71166:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.clo
ud:31929/BLUDB
Done.
```

Out[93]:

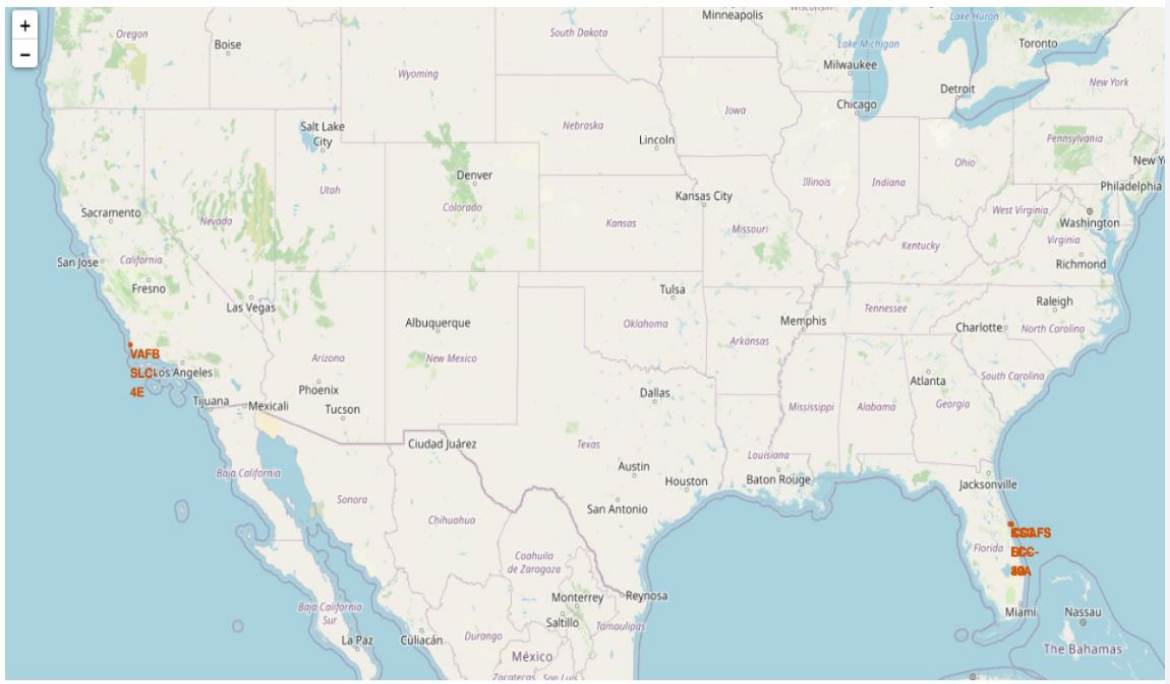
landing_outcome	count_outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Success (ground pad)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	1
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

All Launch Sites' Location on a Global Map

- All launch sites are in close proximity to the equator line
- All launch sites are in very close proximity to the coast
- All launch sites are away from cities or populated areas



Colour-Labelled Markers On Launch Sites

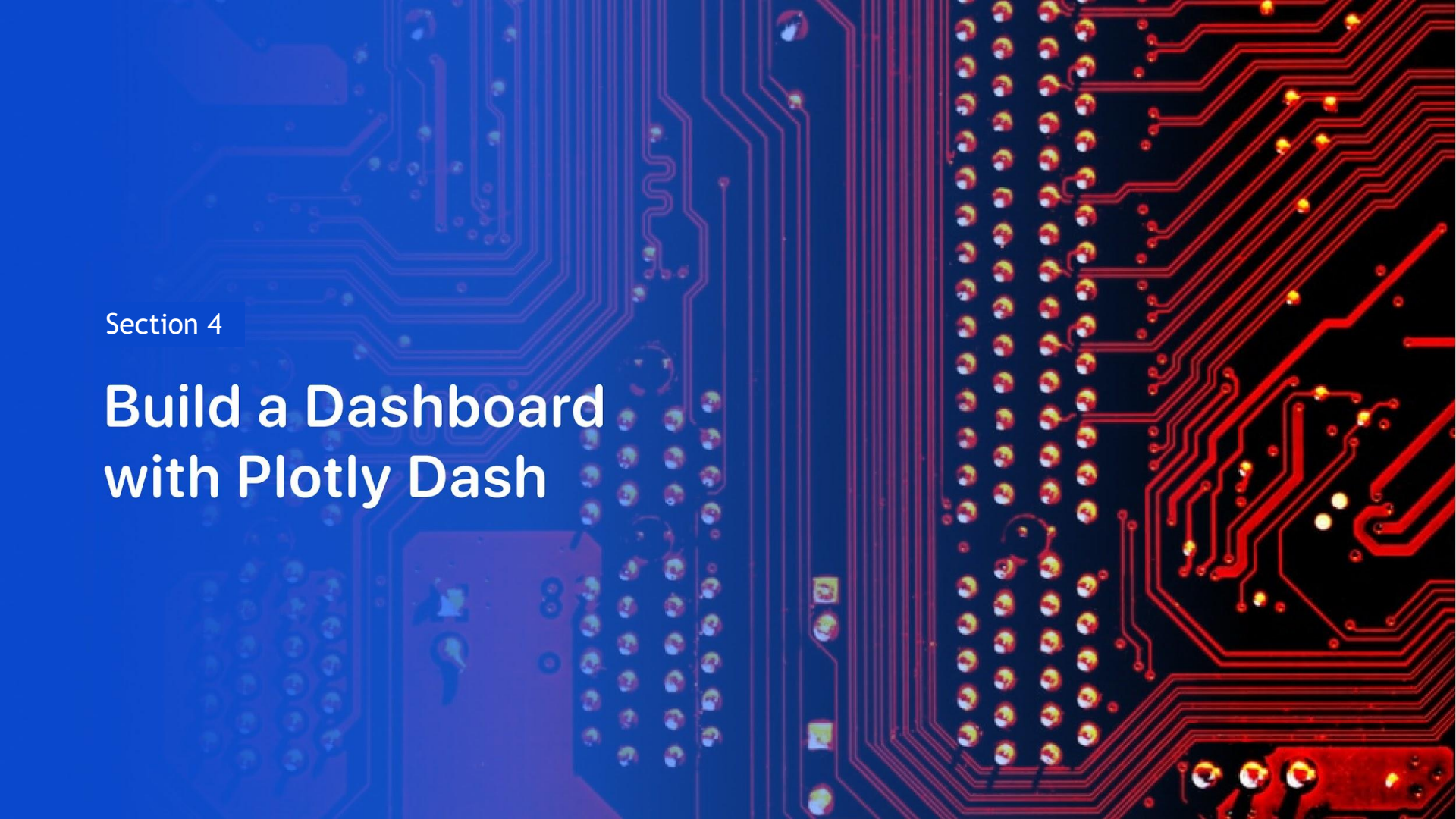
- Colour-labelled markers allow us to easily identify which launch sites have relatively high success rates
- Green Markers are for successful launches while Red ones are for failed launches
- The screenshot is an example of only one launch site

Distance to Closest Coastline from Launch Site

- We have used a line marker to show the distance from the launch site to the closest coastline
- It is very near to the coastline (0.9 kms)
- This example is only for coast line. Please refer to appendix for more

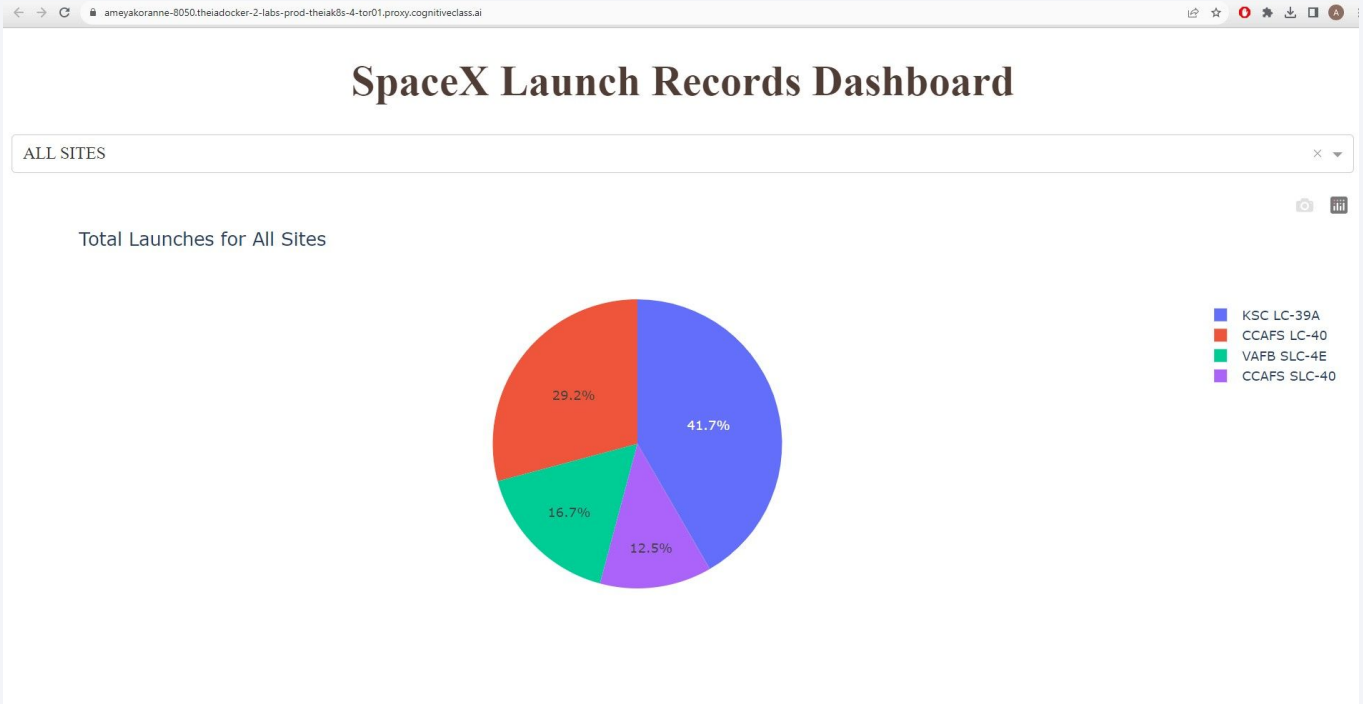

```
launch_site_lat, launch_site_lon = 28.563197, -80.576820
coastline_lat, coastline_lon = 28.56381, -80.5679

distance_coastline = calculate_distance(launch_site_lat, launch_site_lon, coastline_lat, coastline_lon)
print("Distance to closet coastline is", distance_coastline, "Kms.")
```

```
Distance to closet coastline is 0.8740742894031447 Kms.
```


Section 4
Build a Dashboard with Plotly Dash

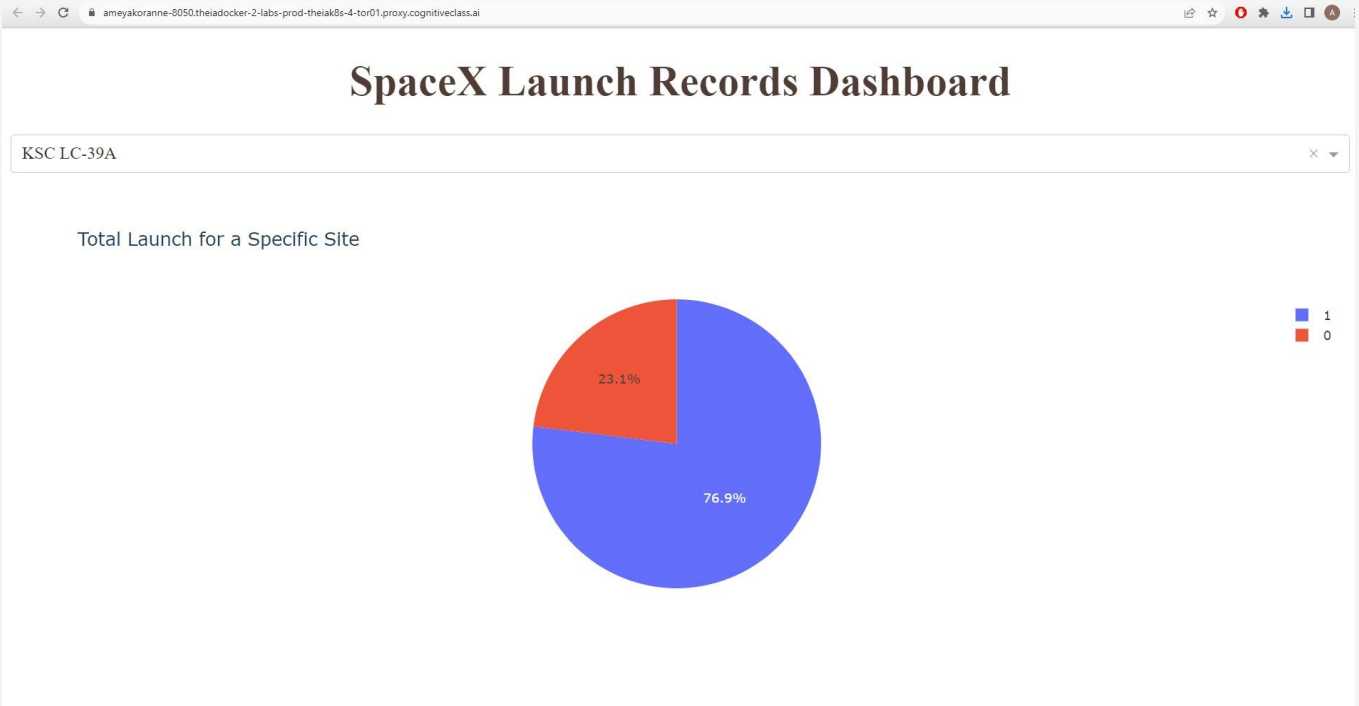
Total Successful Launch Counts for All sites

- Pie Chart showing the launch successful counts for all sites
- We can clearly see KSC LC-39A has the most successful launches



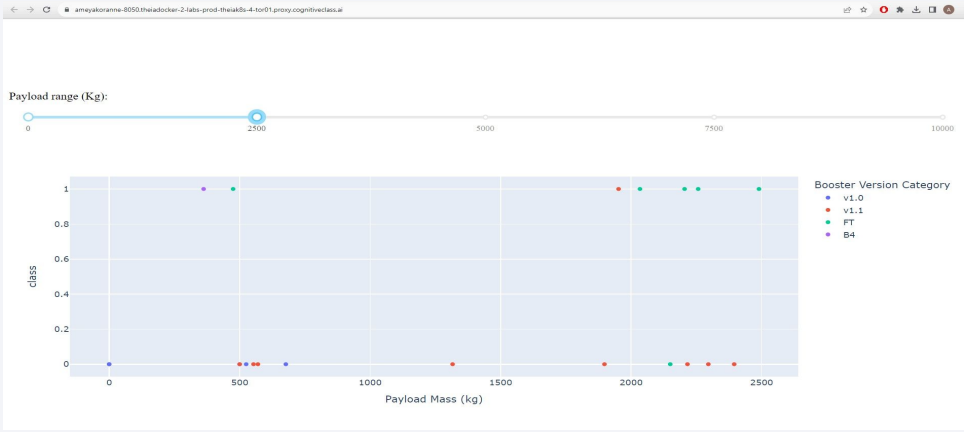
Launch Site with Highest Launch Success Ratio

- Screenshot showing the piechart for the launch site with highest launch success ratio
- KSC LC-39A has a 76.9% success ratio, highest amongst all launch sites



Payload vs. Launch Outcome Ranges

- Showing screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider

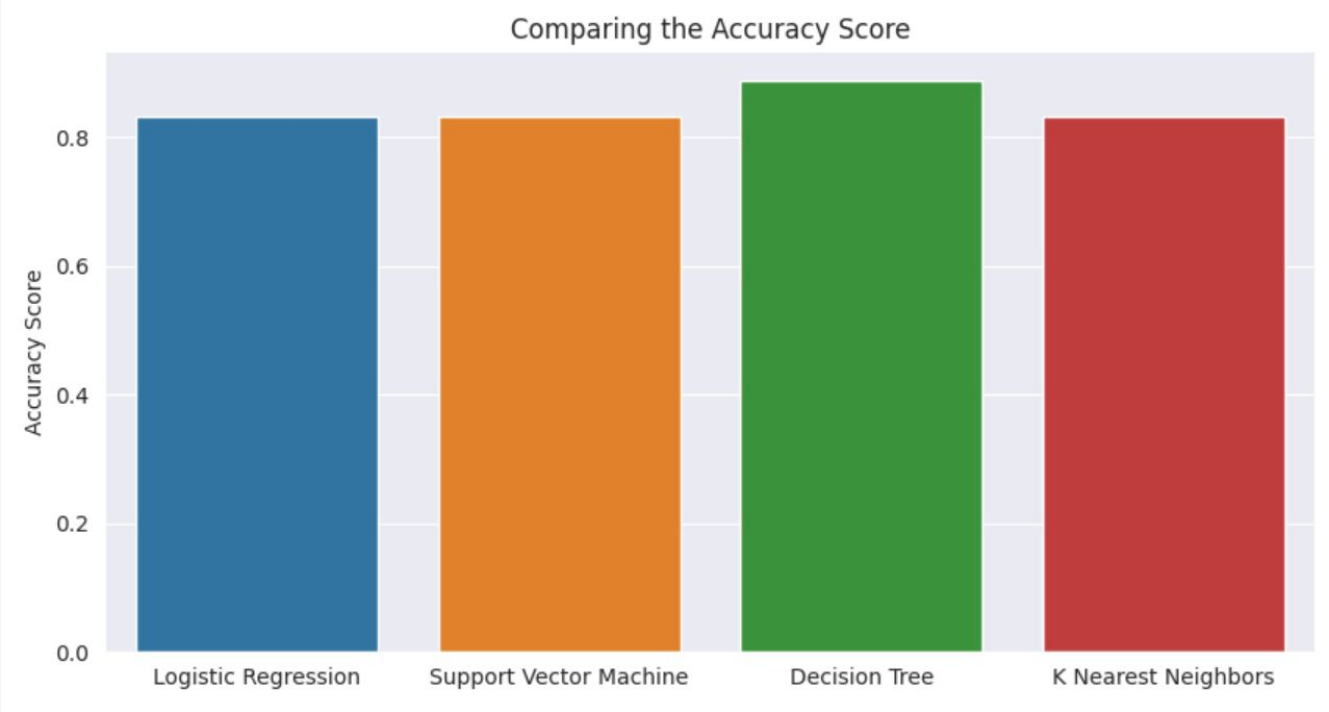


Section 5

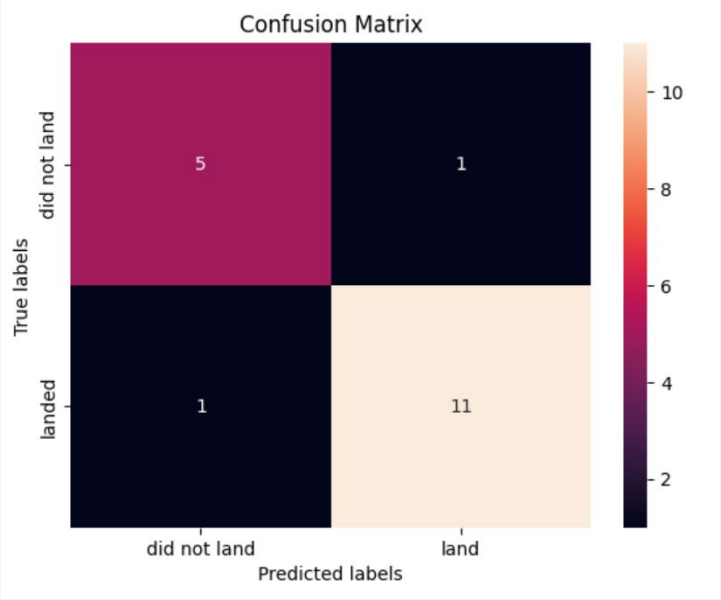
Predictive Analysis (Classification)

Classification Accuracy

- We can easily see that Decision Tree model has the highest classification accuracy (0.88) and the rest are slightly behind at (0.83)



Confusion Matrix



- Showing the confusion matrix of the best performing model - Decision Tree

```
Best model is DecisionTree with a score of 0.8857142857142858
Best params is : {'criterion': 'entropy', 'max_depth': 8, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split':
10, 'splitter': 'best'}
```

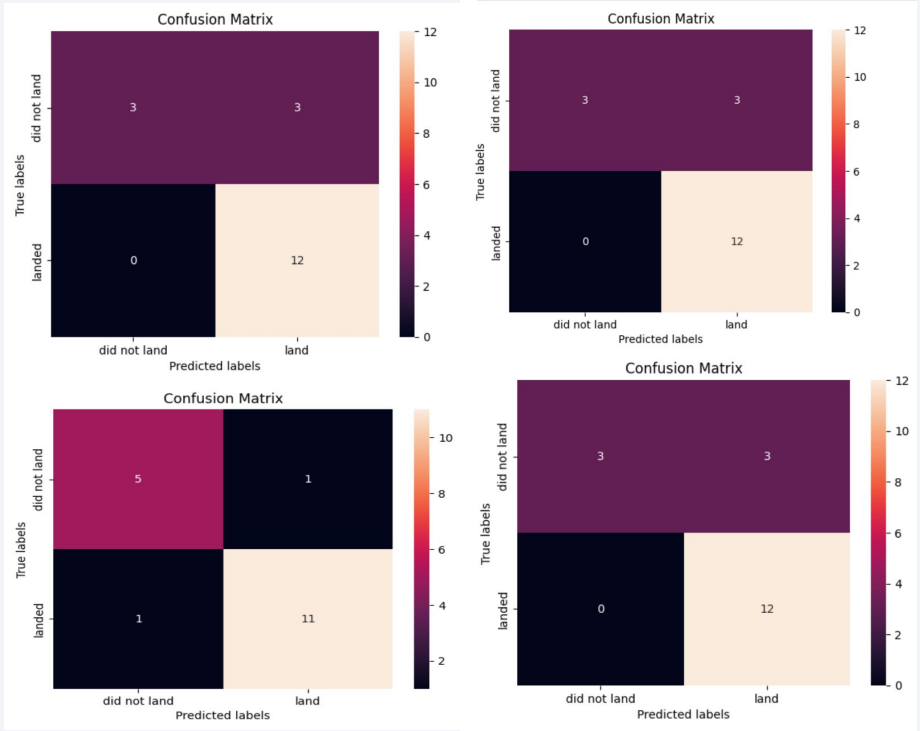
Conclusions

- We can conclude that first stage landings on a ‘drone ship’ and ‘ground pad’ have better success rates
- Payload Mass below 6000 kgs have a higher success rate as compared with heavier payloads
- Classification evaluation metric suggests that Decision Tree is the best algorithm for our dataset
- Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate
- Launch Site KSC LC-39A has the most successful (77%) launch rate amongst all the launch sites
- We can see the upward trend that success rate is growing since 2013

Appendix - Comparison

- A quick comparison of all models - Logistic Regression, SVM, Decision Tree and KNN algorithms
- Sample code to generate Confusion Matrix:

```
yhat = knn_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



Appendix - Code to Compare Best Accuracy and Best Score (BarPlot)

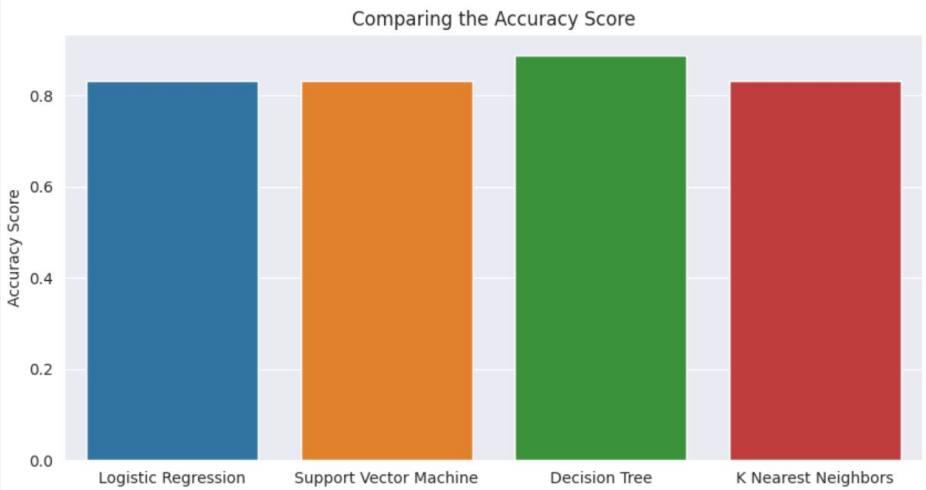
Best Accuracy

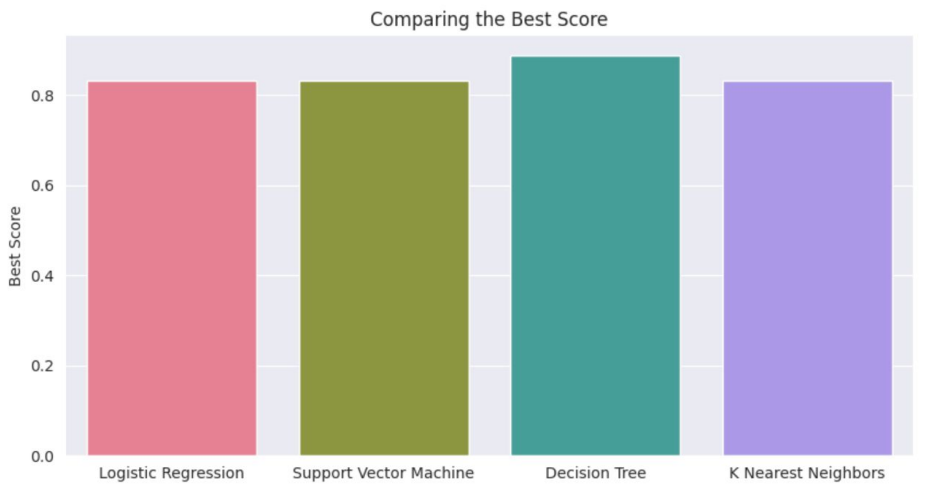
```
sns.set_style("darkgrid")
plt.figure(figsize=(10,5))
sns.barplot(x = algorithms, y = scores, palette = "tab10")
plt.title("Comparing the Accuracy Score")
plt.ylabel("Accuracy Score")
plt.show()
```

Best Score

```
sns.set_style("darkgrid")
plt.figure(figsize=(10,5))
sns.barplot(x = algorithms, y = scores, palette = "husl")
plt.title("Comparing the Best Score")
plt.ylabel("Best Score")
plt.show()
```

Appendix - Best Accuracy vs. Best Score





	Algorithm	Accuracy Score	Best Score
0	Logistic Regression	0.833333	0.846429
1	Support Vector Machine	0.833333	0.848214
2	Decision Tree	0.888889	0.885714
3	K Nearest Neighbors	0.833333	0.848214

- Comparing the results side by side, we learn that Decision Tree is the best performing model for our analysis. It is slightly better than all the other models we used.

Appendix - Payload Mass vs. Class, Best Booster Version

- We learn that success rate is higher when payload mass is lesser than 6000kgs and Booster Version FT has highest success rate

Thank you!